What is the Technology Scope for Building a Geospatial Cyberinfrastructure to Support EarthCube? A white paper for EarthCube revised from a review paper

Chaowei Yang (GMU), Robert Raskin (NASA JPL), Michael Goodchild (UCSB), Mark Gahegan (Auckland)

1. Introduction

A Cyberinfrastructure (CI) is a combination of data resources, network protocols, computing platforms, and computational services that brings people, information, and computational tools together to perform science or other data-rich applications in this information-driven world (NSF 2007 and NSF 2009). Most science domains adopt intrinsic geospatial principles (such as spatial constraints in phenomena evolution) for large amounts of geospatial data processing (such as geospatial analysis, feature relationship calculations, geospatial modeling, geovisualization, and geospatial decision support). Geospatial CI (GCI) refers to CI that utilizes geospatial principles and geospatial information to transform how research, development, and education are conducted within and across science domains (such as the environmental and Earth sciences). GCI is based on recent advancements in geospatial science, information technology, computer networks, sensor networks, Web computing, CI, and e-research/e-science. . GCI provides significant improvements to how the sciences that need geospatial information will advance. The evolution of GCI will produce platforms for geospatial science domains and communities to better conduct research and development and to better collect data, access data, analyze data, model and simulate phenomena, visualize data and information, and produce knowledge. To achieve these transformative objectives, **collaborative research and federated developments** are needed for the building a sustainable CI to support the new Earth sciences, geospatial applications, and education in the 21st century. This white paper is an excerpt of a full review paper about the research, development, education, and other efforts that have contributed to building GCI in terms of its history, objectives, architecture, supporting technologies, functions, application communities, and future research directions

2. History

The history of building an infrastructure to support Earth science research can be traced back to 1800s when USGS and other agencies collaborated to build a national topology base map (Fig.1). The CI term was initialized in 1998 by the White House director and coordinator for national infrastructure, and was widely recognized as a study domain through the NSF (2003) establishment of the Office of CI (OCI). The FGDC NSDI (Nebert 2004), international Digital Earth initiative, and the intergovernmental Group on Earth Observations (GEO) as well as many other geospatial initiatives contribute significantly to the general direction of GCI to build an IT infrastructure for the new Earth science in the 21st century.

 This is a shortened version from a review paper, for original details, please refer to "Yang, C., Raskin, R., Goodchild, M.F., and Gahegan, M., 2010, Geospatial Cyberinfrastructure: Past, Present and Future, Computers, Environment, and Urban Systems, 34(4):264-277."

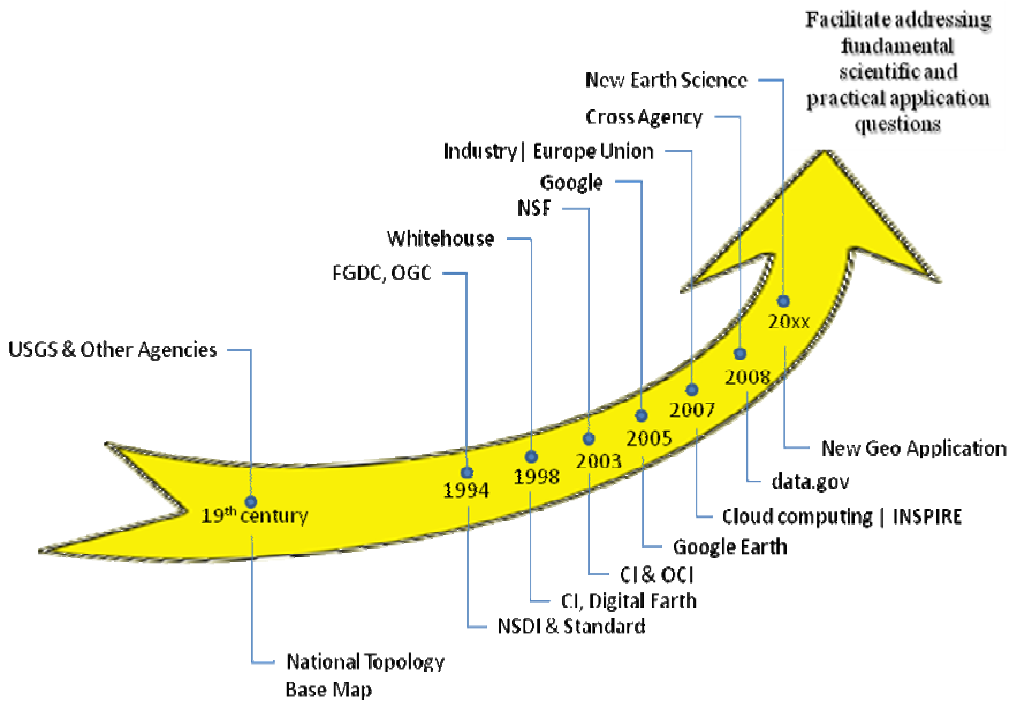

Figure 1. History of Geospatial Cyberinfrastructure

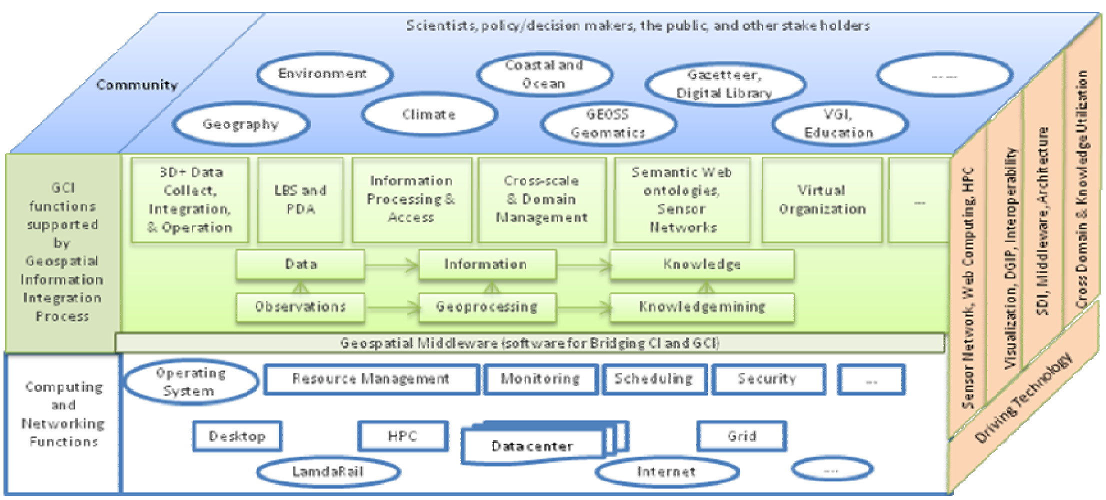

Figure 2. GCI Framework Cube

3. GCI Cube Framework

GCI includes multiple categories of resources within a flexible, scalable, and expandable framework cube (Fig. 2): 1) **Functions** include both generic CI functions (computing, networking, and hardware) and those that are geospatial-specific. 2) The **community** represents the virtual organizations and end user interactions within specific communities including geographic, environmental, Earth, and other science domains. 3) **Enabling**

This is a shortened version from a review paper, for original details, please refer to "Yang, C., Raskin, R., Goodchild, M.F., and Gahegan, M., 2010, Geospatial Cyberinfrastructure: Past, Present and Future, Computers, Environment, and Urban Systems, 34(4):264-277."

technologies provide technological support to invent, mature, and maintain all GCI functions, such as collecting data through observations and collecting and utilizing knowledge through a semantic web.

4. Enabling Technologies

The architecture and integration of GCI benefit from numerous enabling technologies, many of which contributed to the birth of GCI and include, but not limited to:

1) Earth observation and sensor networks provide data collection capabilities to feed petabytes of data into a GCI on a daily basis (Freudinger 2009).
2) The SDI made vast amounts of geospatial information that were collected by government agencies or private companies publicly available and built them into open services to be integrated into customized scientific and practical applications (Nebert 2004).
3) Distributed geospatial information processing (DGIP, Yang et al., 2008) and Web computing (Pierce et al. 2009) handle geospatial information for GCI using dispersed computing resources across platforms.
4) Open and interoperable access technologies (Yang et al. 2007), such as XML/GML, Javascript, and AJAX, enable geospatial data to be published, accessed easily, and adapted to customized applications such as mashups.
5) High-Performance Computing (HPC) provides computing power for GCI users to conduct data and computing-intensive research that cannot be conducted on single computers (e.g., Wang and Liu 2009; Xie et al. 2010).
6) Open-source software and middleware provide a glue function to integrate the components of data, processing, applications, and infrastructure (http://www.nsf.gov/news/special_reports/cyber/middleware.jsp).
7) Solutions to data visualization needs in applied science made it easy to engage end-users to utilize and contribute to a GCI.
8) Cross-domain sharing and collaborations are essential to the ability of a GCI to support and leverage expertise across user communities (MacEachren and Brewer 2004).
9) Knowledge capture and utilization provide for the smart discovery, integration, indexing, collection, and utilization of vast amounts of data, processing components, and other tools available in a GCI (Brodaric, Fox, and McGuinness 2009).
10) System integration architectures (Yang et al., 2008) include Multi-tier/layer organizations (Zhang and Tsou 2009), Mashup and plug-and-play (Bambacus et al. 2007), SOA (Hey and Trefethen (2005), and Workflow chaining (Minsker et al. 2006).

5. GCI Functions

The basic functional components within a GCI provide users access to geospatial data, information, knowledge, and processing tools and include but not limited to:

1) Spatiotemporal data processing (Yang 2000; Yuan and Horns 2005)
2) Data collection & heterogeneous integration (Nittel et al. 2004; Horsburgh 2009)

This is a shortened version from a review paper, for original details, please refer to "Yang, C., Raskin, R., Goodchild, M.F., and Gahegan, M., 2010, Geospatial Cyberinfrastructure: Past, Present and Future, Computers, Environment, and Urban Systems, 34(4):264-277."

3) Data preservation and accessibility (Berman 2008)
4) Supporting the life cycle from data to knowledge (Bolgman et al. 2007)
5) Metadata (such as Henry and Friedlander 2008) for results representation and visualization, services integration, and uniform access to data and information services
6) Virtual Organizations (VO) (Myhill 2008 and Baker 2009),
7) Semantic Web and knowledge sharing,
8) HPC and associated spatial computing,
9) Location Based Service,
10) Cross-scale and domain management.

6. Application Domains and User Communities

Geospatial principles intrinsically reside in almost all scientific domains (Yang et al. 2010) and GCI functions support multiple domains:

1) Geographic sciences
2) Environmental sciences
3) Climate sciences
4) Coastal and ocean studies
5) Broader Earth-system sciences
6) GEOSS societal benefit areas
7) Geomatics
8) Digital libraries
9) Education

7. Discussion and Future Strategies

To achieve the transformative function of GCI, we need to advance at least the following aspects:

1) Studying social heterogeneity to identify geospatial problems encountered by relevant sciences and applications in the context of GCI.
2) Analyzing the information flow from data, information, to knowledge and the processing needed for solving problems.
3) Utilizing the semantic Web to support building knowledge and semantics into the next generation of scientific tools will support smart processing of geospatial metadata, data, information, knowledge, and services for virtual communities and multiple scientific domains (Berners-Lee, Hendler, and Lassila 2001; Hendler 2003).
4) Developing geospatial middleware to provide functional and intermediate services and support service evolution to stakeholders.
5) Advancing citizen sciences or public-based sciences will reflect the fact that cyberspace is open to the public and citizen participation will play vital roles.
6) Advancing GCI to geospatial cloud computing will implement transparent and opaque platforms for addressing fundamental science questions and application

This is a shortened version from a review paper, for original details, please refer to "Yang, C., Raskin, R., Goodchild, M.F., and Gahegan, M., 2010, Geospatial Cyberinfrastructure: Past, Present and Future, Computers, Environment, and Urban Systems, 34(4):264-277."

problems through advancements in information technology and computing sciences.

7) Developing a research agenda that can address the needs of GCIs through effective federation and collaboration across communities, such as government agencies, non-government organizations, industries, academia, and the public.

The seven aspects described will benefit from and drive the enabling technologies for future GCIs (Table 1) and provide improved functions for GCI application domains (Table 2).

Table 1 Connections between enabling technologies and future research needs

	7.1 social	7.2 information	7.3 semantics	7.4 middleware	7.5 citizen science	7.6 cloud computing	7.7 research agenda
4.1 observation		x	x				x
4.2 SDI	x	x	x		x		x
4.3 DGIP		x	x	x		x	x
4.4 Web	x	x	x	x	x		x
4.5 interoperability	x	x	x	x	x	x	x
4.6 HPC		x	x	x		x	x
4.7 middleware		x	x			x	x
4.8 representation	x	x	x		x		x
4.9 cross domain	x		x		x		x
4.10 knowledge	x	x	x	x	x	x	x
4.11 architecture	x	x	x			x	x

Table 2 Connections between GCI functions and future research needs

	7.1 social	7.2 information	7.3 semantics	7.4 middleware	7.5 citizen science	7.6 cloud computing	7.7 research agenda
5.1 data processing		x	x	x			x
5.2 data integration	x	x	x	x		x	x
5.3 preservation & accessibility	x	x	x	x		x	x
5.4 data life cycle	x	x	x	x	x	x	x
5.5 VO	x	x	x	x	x	x	x
5.6 semantics	x	x	x		x	x	x
5.7 HPC		x	x	x		x	x
5.8 LBS	x	x	x		x	x	x
5.9 cross domain	x		x		x	x	x

This is a shortened version from a review paper, for original details, please refer to "Yang, C., Raskin, R., Goodchild, M.F., and Gahegan, M., 2010, Geospatial Cyberinfrastructure: Past, Present and Future, Computers, Environment, and Urban Systems, 34(4):264-277."

References

1. Berman, F., 2008. Got data? A guide to data preservation in the information age, *Communications of the ACM*, 51(12): 50-56.

2. Berners-Lee T., Hendler J., and Lassila O., 2001. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American*, May 17, 2001.

3. Borgman C.L., Wallis J.C., Mayernik M.S., Pepe A., 2007. Drowning in data: Digital library architecture to support scientific use of embedded sensor networks, In *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries, JCDL 2007: Building and Sustaining the Digital Environment*, Jun 18-23 2007, Vancouver, BC, p 269-277

4. Brodaric B., Fox P., McGuinness D. L., 2009. Geoscience knowledge representation in cyberinfrastructure, *Computers and Geosciences*, 35(4):697-699.

5. Nebert D., 2004. GSDI Cook Book Version 2.0 171p.

6. Freudinger, L. C., 2009. Cyberinfrastructure for Airborne Sensor Webs, *NASA Dryden Flight Research Center Report: DFRC-927*, 5p.

7. Goodchild M.F., 2008. Commentary: whither VGI? *GeoJournal*, 72: 239–244.

8. Hendler J., 2003. Enhanced: Science and the Semantic Web, *Science*, 299(5606): 520-521.

9. Henry C.J., Friedlander A., 2008. Cataloging hidden special collections and archives: building a new research environment, *Joint Conference on Digital Libraries (JCDL 2008)*, June 15-19, 2008, Austin, TX, p.439.

10. Hey T., and Trefethen A.E., 2005. Cyberinfrastructure for e-Science, *Science*, 308(5723):817-821.

11. Horsburgh J. S., Tarboton D. G., Piasecki M., Maidment D. R., Zaslavsky I., Valentine D., Whitenack T., 2009. An integrated system for publishing environmental observations data, *Environmental Modelling & Software*, 24(8):879-888.

12. MacEachren, A. M., and Brewer, I., 2004. Developing a conceptual framework for visually-enabled geocollaboration, *International Journal of Geographical Information Science*, 18(1):1-34.

13. Minsker B., Myers J., Marikos M., Wentling T., Downey S., Liu Y., Bajcsy P., Kooper R., Marini L., Contractor N., Green H.D., Futrelle J., 2006. NCSA environmental cyberinfrastructure demonstration project: Creating cyberenvironments for environmental engineering and hydrological science

communities, *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, SC'06, 2006*, Nov 11-17 2006, Tampa, FL, United States, p 1188622.

14. Myhill W. N., Cogburn D. L., Samant D., Addom B. K., Blanck P., 2008. Developing Accessible Cyberinfrastructure-Enabled Knowledge Communities in the National Disability Community: Theory, Practice, and Policy, *Assistive Technology*, 20(3):157-174.

15. Nittel S., Stefanidis A., Cruz I., Egenhofer M., Goldin D., Howard A., Labrinidis A., Madden S., Voisard A., Worboys M., 2004. Report from the first workshop on geo sensor networks, *ACM SIGMOD Record*, 33(1):141-144.

16. NSF, 2003, *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. NSF Panel reports. 84pp.

17. NSF, 2007. *Cyberinfrastructure Vision for 21st Century Discovery*, 64pp.

18. NSF, 2009. *NSF-Supported Research Infrastructure: Enabling Discovery, Innovation and Learning*, NSF, NSF-09-13, 148p.

19. Pierce M. E., Fox G. C., Choi J. Y., Guo Z., Gao X., Ma Y., 2009. Using Web 2.0 for scientific applications and scientific communities, *Concurrency Computation Practice and Experience*, 21(5):583-603.

20. Wang S. W., Liu Y., 2009. TeraGrid GIScience Gateway: Bridging cyberinfrastructure and GIScience, *International Journal of Geographic Information Science*, 23(5):631-656.

21. Xie J., Yang C., Zhou B, and Huang Q., 2010. High-performance computing for the simulation of dust storms, *CEUS*, 4(34): in press.

22. Yang, C., 2000, Theory, Techniques and Implementation Methods of WebGIS, Unpublished Ph.D. dissertation (Beijing: Peking University) (in Chinese).

23. Yang C., Li W., Xie J., and Zhou B., 2008. Distributed geospatial information processing: sharing earth science information to support Digital Earth, *International Journal of Digital Earth*. 1(3): 259-278.

24. Yang P., Evans J., Cole M., Alameh N., Marley S., and Bambacus M., 2007. The Emerging Concepts and Applications of the Spatial Web Portal, *PE&RS*, 73(6):691-698.

25. Zhang T., Tsou M. H., 2009. Developing a grid-enabled spatial Web portal for Internet GIServices and geospatial cyberinfrastructure, *International Journal of Geographic Information Science*, 23(5):605-630.